Available online at www.sciencedirect.com
SCIENCE DIRECT®
Genomics 84 (2004) 211–217
GENOMICS
www.elsevier.com/locate/ygeno

A computational approach to measuring coherence of gene expression in pathways

Howard H. Yang, Ying Hu, Kenneth H. Buetow, and Maxwell P. Lee*

Laboratory of Population Genetics, National Cancer Institute, 41 Library Drive, Bethesda, MD 20892, USA

Received 15 May 2003; accepted 23 January 2004

Available online 14 March 2004

Abstract

This study uses a computational approach to analyze coherence of expression of genes in pathways. Microarray data were analyzed with respect to coherent gene expression in a group of genes defined as a pathway in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Our hypothesis is that genes in the same pathway are more likely to be coordinately regulated than a randomly selected gene set. A correlation coefficient for each pair of genes in a pathway was estimated based on gene expression in normal or tumor samples, and statistically significant correlation coefficients were identified. The coherence indicator was defined as the ratio of the number of gene pairs in the pathway whose correlation coefficients are significant, divided by the total number of gene pairs in the pathway. We defined all genes that appeared in the KEGG pathways as a reference gene set. Our analysis indicated that the mean coherence indicator of pathways is significantly larger than the mean coherence indicator of random gene sets drawn from the reference gene set. Thus, the result supports our hypothesis. The significance of each individual pathway of n genes was evaluated by comparing its coherence indicator with coherence indicators of 1000 random permutation sets of n genes chosen from the reference gene set. We analyzed three data sets: two Affymetrix microarrays and one cDNA microarray. For each of the three data sets, statistically significant pathways were identified among all KEGG pathways. Seven of 96 pathways had a significant coherence indicator in normal tissue and 14 of 96 pathways had a significant coherence indicator in tumor tissue in all three data sets. The increase in the number of pathways with significant coherence indicators may reflect the fact that tumor cells have a higher rate of metabolism than normal cells. Five pathways involved in oxidative phosphorylation, ATP synthesis, protein synthesis, or RNA synthesis were coherent in both normal and tumor tissue, demonstrating that these are essential genes, a high level of expression of which is required regardless of cell type.
Published by Elsevier Inc.

The recent completion of the human genome sequence has been accompanied by increasing interest in high-throughput genomics-based global expression technologies. These methods facilitate studies of the cellular transcriptome or proteome and the networks of interactions between the genome, the transcriptome, and the proteome. Affymetrix DNA oligonucleotide microarray and cDNA microarray are the two major platforms for genome-wide analysis of transcription and gene regulatory networks. These technologies have the capacity to measure simultaneously transcription of all of the estimated 30,000 genes in the human genome.

Many experimental studies have successfully employed microarray technology. For example, microarray analysis has proven to be a powerful approach for classifying cancers. Two clinically distinct types of diffuse large B cell lymphoma were identified by clustering samples based on their gene expression profiles. This information facilitates increased precision of diagnosis, staging, and prediction of cancer prognosis [1]. The clustering methods developed for microarray data analysis include dendrogram [2], *K*-means [3], self-organizing map [4], support vector machine [5], and neural network [6]. Microarray data have also been used to demonstrate coordinate regulation over time of genes involved in the metabolic shift from fermentation to respiration [7] and cell cycle progression in *Saccharomyces cerevisiae* [8]. Coregulation of ligand–receptor gene pairs has also been studied using microarray technology [9]. Despite the successes in clustering samples and genes based on microarray gene expression data, the

* Corresponding author. Fax: +1-301-435-8963.
E-mail address: leemax@mail.nih.gov (M.P. Lee).

0888-7543/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.ygeno.2004.01.007

existing pathway knowledge has seldom been applied in these analyses and no single index has been provided to quantify the expression of genes in each gene set identified by these approaches. Genes have usually been divided into nonoverlapping subsets in these approaches, while Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and other biological networks often share some common genes. This study reports a computational approach to measuring coherence of gene expression for genes in a pathway. We developed a single index to quantify coordinate regulation of all genes in each pathway. Pearson's correlation coefficient was used to assess the correlation in expression of all gene pairs in a pathway. The coherence indicator was defined as the ratio of the number of gene pairs in a pathway whose correlation coefficients are significant, divided by the total number of gene pairs in the pathway. Coherence indicators were computed for each of the 96 KEGG pathways in normal and tumor samples using three different data sets, and the significance of the coherence indicators was assessed.

Results

This study is based on the hypothesis that the expression of two genes in the same pathway is more likely to be correlated than that of two genes in different pathways. This hypothesis predicts that a group of genes in one pathway is more likely to be coordinately regulated than a random set of genes. This hypothesis was tested in the following manner. First, Pearson correlation coefficient was calculated for all gene pairs in each of the 96 pathways described in the KEGG system, and significantly correlated gene pairs were identified. Fig. 1 shows the results for genes in the pathway for fructose and mannose metabolism. Results are presented as scatter plots of

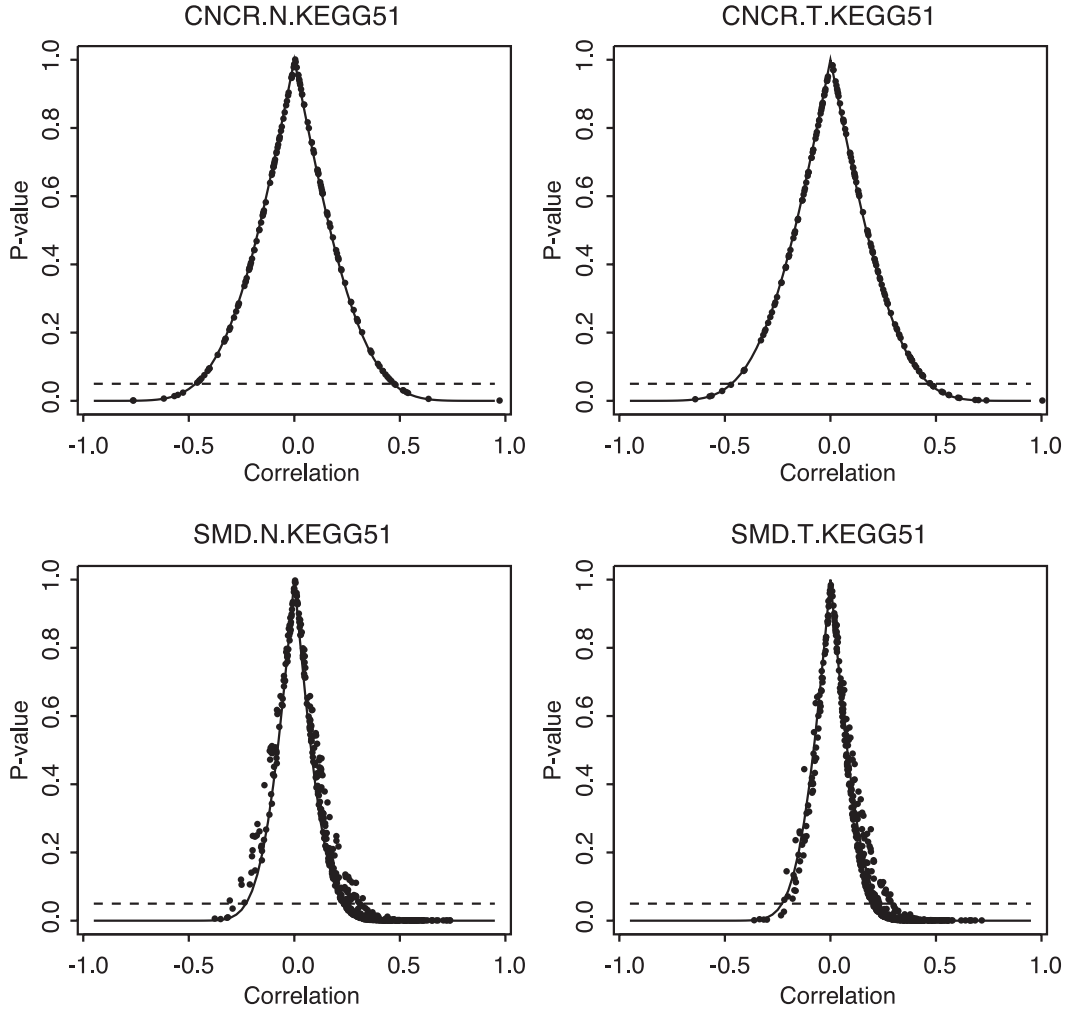

Fig. 1. Scatter plot of correlation coefficient vs p value for normal or tumor samples in CNCR and SMD data sets. Gene pairs are from the fructose and mannose metabolism pathway. Normal samples are shown on the left and tumor samples are shown on the right. CNCR data are shown on the top and SMD data are shown on the bottom. The dashed line is the 5% significance level. The effective sample size for each gene pair is 18 or fewer in the CNCR and 74 or fewer in the SMD data set. The solid line shows the two-tailed p value of t statistics for sample size of 18 (top) and 74 (bottom).

correlation coefficient vs p value for normal and tumor samples in two data sets. The solid line in each graph shows the two-tailed p values of the t statistics with $(m - 2)$ degrees of freedom when the effective sample size is m. The t statistic was computed by $T = \text{sqrt}(m - 2) \times R/\text{sqrt}(1 - R^2)$ for a correlation coefficient R ("sqrt" designates "square root"). Second, a coherence indicator was defined as the ratio of the number of gene pairs in a pathway whose correlation coefficient was significant ($p < 0.05$), divided by the total number of gene pairs in the pathway. Third, we defined all genes that appeared in the KEGG pathways as a reference gene set. For each pathway, 10 random gene sets of the same size as the pathway were randomly selected from the reference gene set. Coherence indicators were computed for the random gene sets derived from gene expression data of normal and tumor samples separately. Fourth, distributions of coherence indicators from all pathways were compared between pathway and random gene sets (Fig. 2). Based on the normal sample, the coherence indicators of the pathway gene set are higher than those of the random gene set (Fig. 2A). The mean coherence indicators are 0.61 and 0.53 for pathway and random gene set, respectively, which is highly significant ($p = 1.9 \times 10^{-12}$, t test). The results support our hypothesis. Interestingly, there was no significant difference in the mean coherence indicator between pathway and random gene set in tumors (Fig. 2B). We then went on to evaluate the distributions of coherence indicators for pathways between normal and tumor samples (Fig. 2C). We found that gene expression in tumor had a higher mean coherence indicator than the normal sample ($p = 0.019$, t test). The increase in the coherence of gene expression is not unique to pathway. Coherence in the random gene set was also increased in the tumor sample in comparison to the normal sample (Fig. 2D). The increased coherence indicator in tumor was mostly due to increased gene expression in tumor (see Discussion).

The statistical significance of an individual coherence indicator in a pathway was evaluated by comparing the coherence indicator of the pathway to the coherence

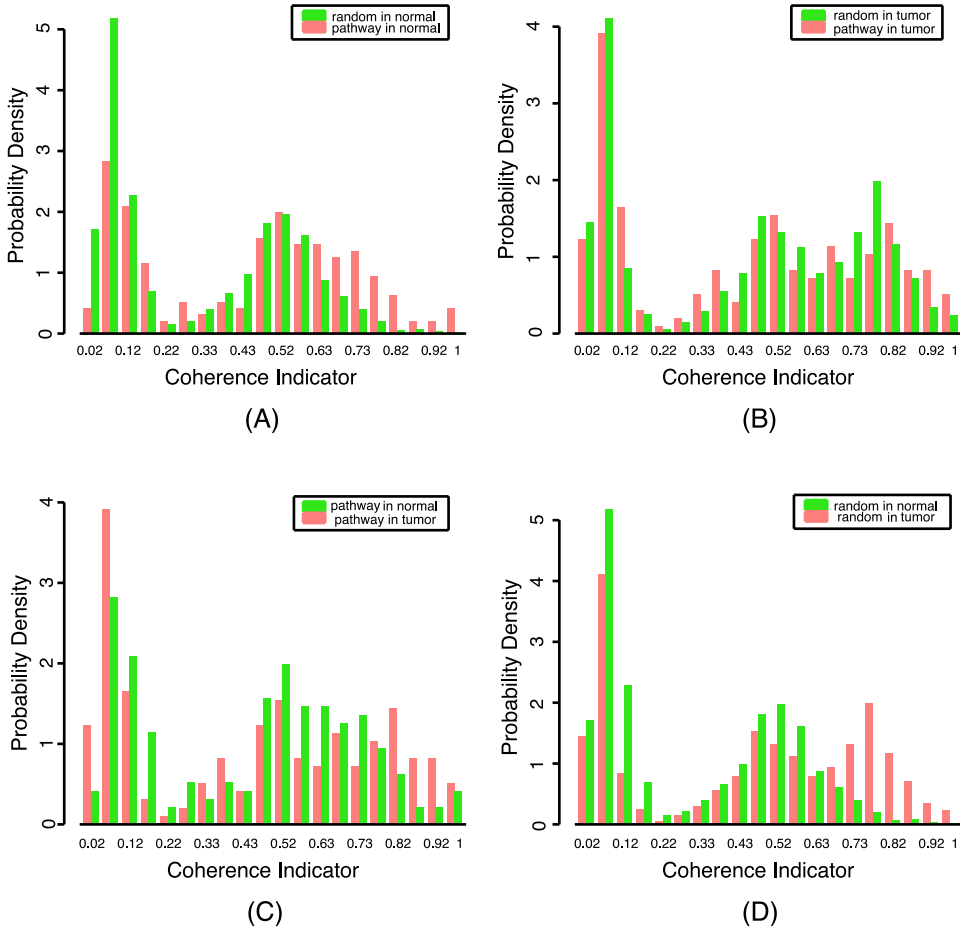

Fig. 2. Comparison of coherence indicator distributions. (A) Pathway vs random gene set in the normal sample. Considering coherence indicators above a threshold of 0.25, the mean coherence indicators (MCIs) for pathways and random gene sets are 0.61 and 0.53, respectively. The t test shows that the former is significantly higher than the latter ($p = 1.9 \times 10^{-12}$). (B) Pathway vs random gene set in the tumor sample. No significant difference between the MCIs of pathway (0.65) and random gene set (0.64) was observed. (C) Tumor vs normal in pathway. The MCI in the tumor sample (0.65) is higher than that in the normal sample (0.61 with $p = 0.02$ in t test). (D) Tumor vs normal in random gene set. The MCI in tumor (0.64) is significantly higher than that in normal (0.53 with $p = 0$ in t test).

indicators of 1000 random gene sets (see Methods). Random gene sets were selected from the reference gene set using a random permutation method, and a coherence indicator was computed for each random gene set. The significance of the coherence indicator for a pathway was evaluated against the background of the coherence indicators of the random gene sets. A pathway was considered to be coherent if its p value was less than 0.05. Although the relation between coherence indicator and p value was unknown, higher coherence indicators tended to have smaller p values (Fig. 3).

The pathways with significant coherence indicators are summarized in Table 1. At least 20% of 96 pathways were coherent in at least one data subset; 7 pathways were coherent in normal samples in three data sets, and 14 pathways were coherent in tumor samples in three data sets (Table 1). A greater number of pathways were also coherent in tumor than in normal samples in two data sets jointly evaluated (Table 1). This is consistent with the global distribution of coherence indicators (Fig. 2D). In addition, 11 pathways were coherent only in tumor samples, but no pathways were coherent only in normal samples (Table 2). Eight pathways were coherent in normal and tumor samples in two data sets jointly evaluated and 5 pathways were coherent in both normal and tumor samples in all three data sets (Table 2). Tumor cells may have more coherent pathways because they have a higher rate of metabolism than normal cells.

A schematic diagram used to display significantly correlated gene pairs in pyrimidine biosynthesis is shown in Fig. 4.

Discussion

This paper presents a computational approach to analyzing coherence of gene expression in pathways using microarray gene expression data. Coherence indicators were

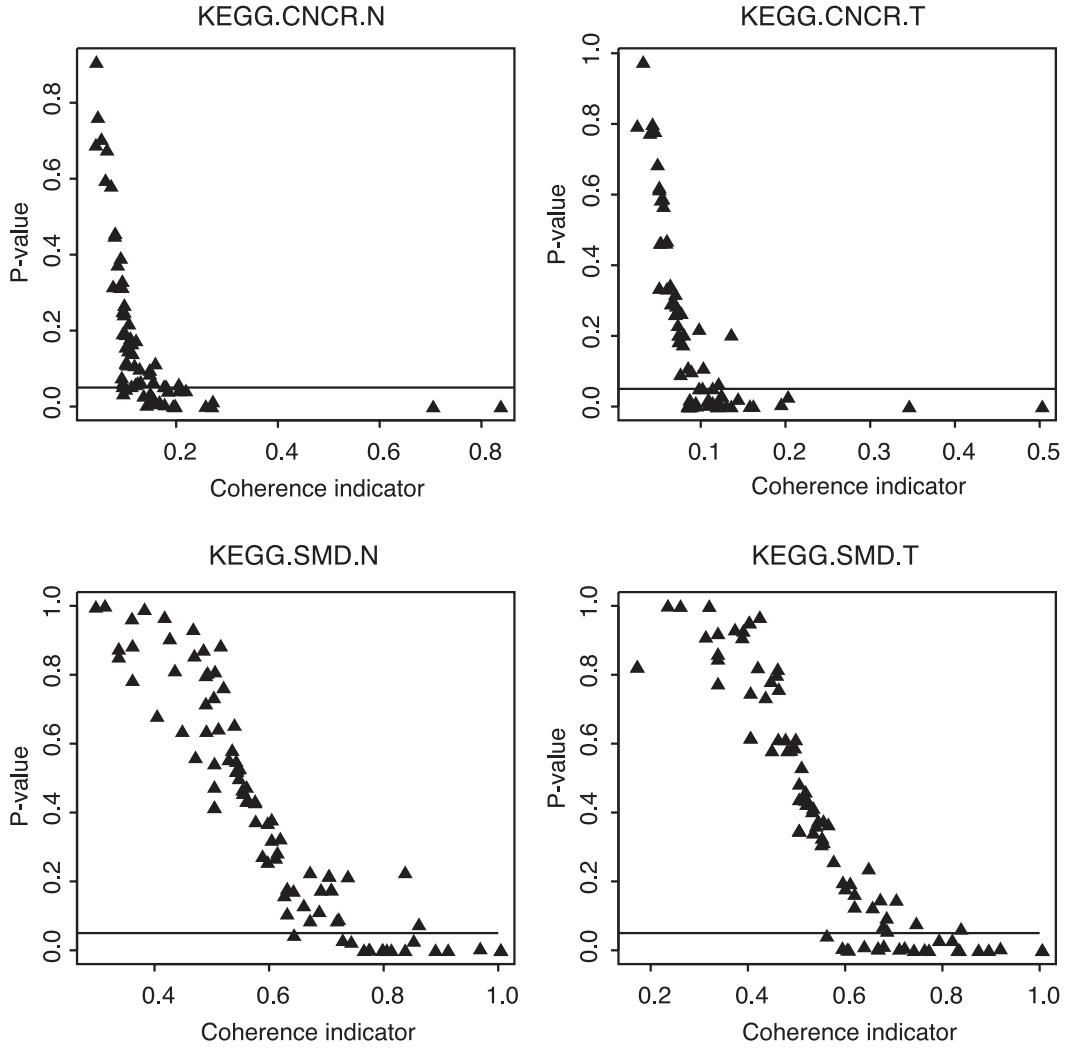

Fig. 3. Coherence indicator vs p value for gene pairs in normal (left) or tumor (right) samples from CNCR (top) and SMD (bottom) data sets. A higher coherence indicator tends to have a smaller p value. The horizontal line indicates $p = 0.05$. Coherence indicators with p values below this line were considered significant.

Table 1
Summary of coherent pathways

Data set	Type	Number of significant pathways
I2000	N	41
I2000	T	53
CNCR	N	21
CNCR	T	29
SMD	N	20
SMD	T	26
CNCR.SMD	N	8
CNCR.SMD	T	15
I2000.CNCR	N	10
I2000.CNCR	T	24
I2000.SMD	N	13
I2000.SMD	T	23
I2000.CNCR.SMD	N	7
I2000.CNCR.SMD	T	14

Data sets are indicated by I2000, CNCR, or SMD. N, normal; T, tumor sample.

estimated in 96 pathways in tumor and normal samples in three microarray data sets. Our initial motivation of this study was to test the hypothesis that genes in the same pathway are more likely to be coordinately regulated than a randomly selected gene set. Our analysis of the distribution of coherence indicators, the ratio of significant correlations among gene pairs in a pathway, clearly supports our hypothesis (Fig. 2A). Thus, coherence indicator is a reliable index for a set of genes sharing a common biological function. In addition, we find that there are more coherent pathways in tumor than in normal samples, including pathways involved in metabolic processes such as synthesis of ATP, protein, and RNA (Fig. 2C, Tables 1 and 2). The increase in the mean coherence indicator in tumor was also observed in the random gene set (Fig. 2D). Upon reexamination of data and analysis, we found that variances in gene expression across samples were smaller in normal than in tumor samples, which were largely caused by low expression of some KEGG genes in normal samples. As explained under Methods, computation of the Pearson coefficient depends on nonzero variance for both genes in the gene pair. This may account for higher mean coherence indicator in tumor than in normal samples. This also explains why the mean coherence indicators were similar between pathway and random gene sets in tumors (Fig. 2B), because the coherence indicators were higher in both pathway and random gene sets in tumors. This is consistent with the fact that tumor cells have a higher rate of metabolism than normal cells. Five pathways were coherent in normal and tumor samples in all three data sets. These pathways include oxidative phosphorylation, ribosomes, and transcription factors, indicating that coherence of these pathways is required for all cells. These experiments demonstrate that biological information can be extracted from microarray data sets using computational analysis and this information can enhance our understanding of important biological processes.

Methods

Gene expression data

The following microarray data sets were analyzed: two Affymetrix oligonucleotide array data sets [10,11] named I2000 and CNCR and one cDNA microarray data set named SMD [12]. There are 22 pairs of colon cancer and matched normal samples in I2000 and 18 pairs of colon cancer and matched normal samples in CNCR. The SMD data set has 180 samples, including 74 pairs of liver cancer and matched normal samples. Each data set was divided into normal and tumor subsets.

I2000 has expression data for 2000 genes selected in [10]. Expression intensity was calculated from the mean of PM−MM intensities.

CNCR has expression data for 7464 genes. Gene expression level for each probe set is the average intensity difference between the PM−MM probe pairs. Following the procedure described in [11], gene expression levels less than 10 were adjusted to 10 in data preprocessing.

SMD has expression data for 42,675 genes, each of which is identified by an accession number. Ninety percent of them can be mapped to Locuslink. Forty-seven percent of the records had missing values in more than 99% of

Table 2
Coherence in normal and tumor samples

Coherence	Data set	Pathway description
In tumor only	I2000.CNCR	Fructose and mannose metabolism
	I2000.CNCR	Sterol biosynthesis
	I2000.CNCR	Urea cycle and metabolism of amino groups
	I2000.CNCR	Pyrimidine metabolism
	I2000.CNCR	Arginine and proline metabolism
	I2000.SMD	Glycoprotein degradation
	CNCR.SMD	Ubiquinone biosynthesis
	CNCR.SMD	Inositol phosphate metabolism
	CNCR.SMD	Sphingoglycolipid metabolism
	CNCR.SMD	Nicotinate and nicotinamide metabolism
	I2000.CNCR.SMD	Apoptosis
In normal only		(None)
In both tumor and normal	I2000.CNCR	Starch and sucrose metabolism
	I2000.SMD	Valine, leucine, and isoleucine degradation
	I2000.SMD	Lysine biosynthesis
	I2000.SMD	Propanoate metabolism
	I2000.SMD	Butanoate metabolism
	I2000.SMD	Protein export
	CNCR.SMD	Photosynthesis
	CNCR.SMD	Aminoacyl-tRNA biosynthesis
	I2000.CNCR.SMD	Oxidative phosphorylation
	I2000.CNCR.SMD	ATP synthesis
	I2000.CNCR.SMD	Ribosome
	I2000.CNCR.SMD	Transcription factors
	I2000.CNCR.SMD	Proteasome

Distribution and description of pathways that were coherent in two or more data sets in tumor and normal samples.

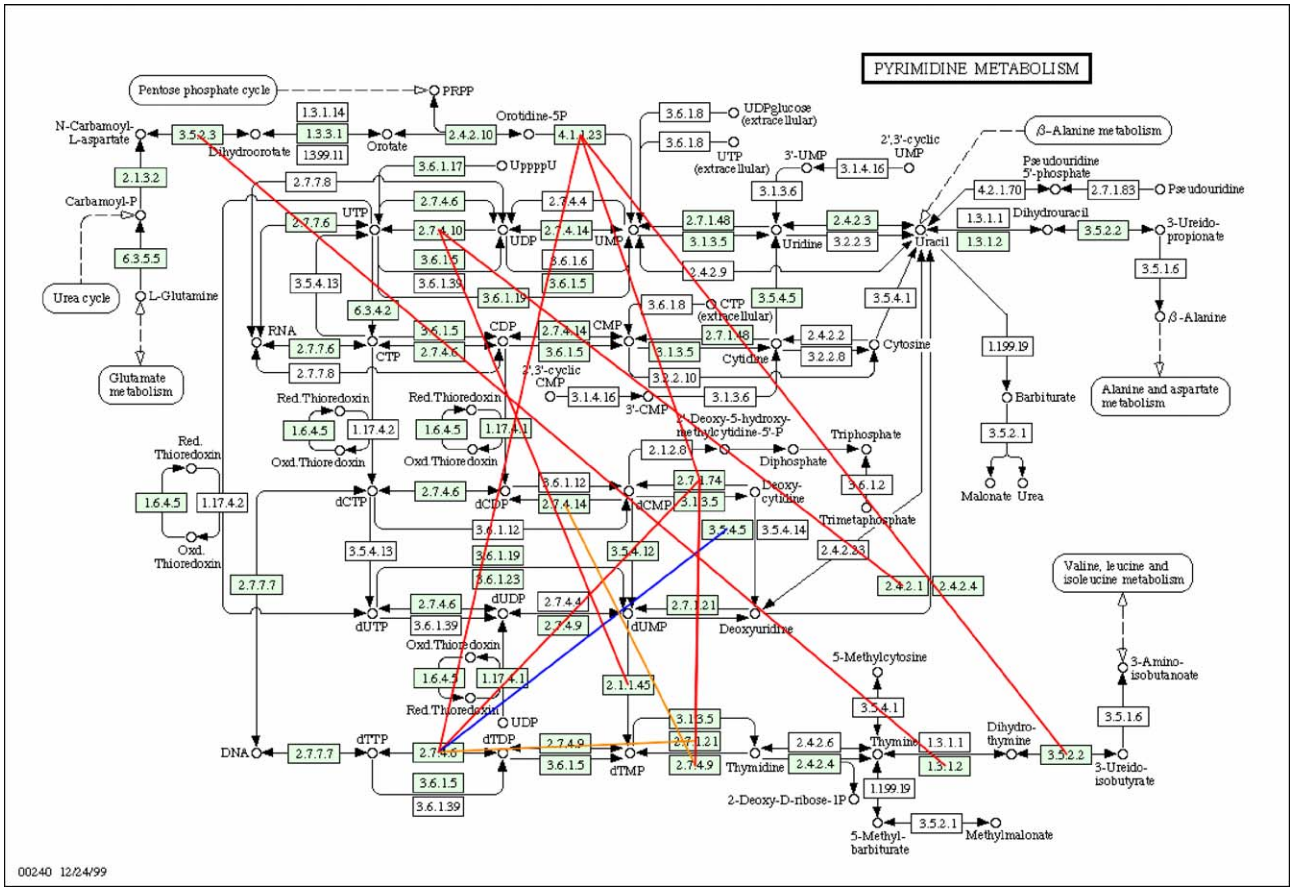

Fig. 4. Diagram of significantly correlated gene pairs in pyrimidine metabolism. Pyrimidine pathway diagram is from http://www.genome.ad.jp/kegg/. Correlation coefficients for gene pairs shown in this pathway were estimated from the CNCR data set. Significantly correlated pairs of genes are connected with a red or blue line if positively or negatively correlated, respectively. Only correlations (red and blue lines) estimated from the tumor samples are shown. Orange lines connect gene pairs with significant positive correlation in both tumor and normal samples in the CNCR data set.

samples and were not included in the analysis. The remaining data set included 22,618 genes. As recommended in Ref. [12], statistics of channel measurement were used to define gene expression according to the following procedure: CH_{1I} and CH_{2I} were defined as the mean intensities in the signal channel and reference channel, respectively, and CH_{1B} and CH_{2B} were defined as the medians of the background in the signal and reference channels, respectively. Gene expression was defined as the ratio $(CH_{1I} - CH_{1B})/(CH_{2I} - CH_{2B})$, if both $(CH_{1I} > 1.5\ CH_{1B})$ and $(CH_{2I} > 1.5\ CH_{2B})$ were true.

Genes selected for analysis belong to pathways defined in the KEGG database (http://www.genome.ad.jp/kegg/). The KEGG database has annotated information on 103 pathways and 1378 genes. The three data sets I2000, CNCR, and SMD contain 330, 809, and 907 genes, respectively, that are present in KEGG pathways.

Coherence indicator

To compute the coherence indicator of a pathway, we identified a pair of genes that satisfy the following two criteria: both genes are included in at least one data set and both genes belong to the same pathway. If the number of the genes in a pathway is n, the number of gene pairs is $n(n-1)/2$. Given a normal or tumor data set of m samples, for each pair of genes in the pathway we evaluated the correlation and its significance in the following steps. In addition to the computation of correlation coefficient and the evaluation of its significance, several conditions were checked in this process. Let $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$ be two sets of gene expression data. Vectors x and y may contain missing values. We first found the effective sample size for x and y that is the number of samples in which x_i and y_i are both present. We then estimated the correlation coefficient for x and y if

(1) the variance Var(x) and the variance Var(y) are both greater than 10^{-10} and
(2) the effective sample size is greater than 2.

If the effective sample size is m, the correlation cannot be tested when $m < 3$ because the number of degrees of freedom $(m - 2)$ must be greater than zero to compute the p value of the correlation. Correlation coefficients with p values <0.05 were considered significant. Gene pairs

whose accession number mapped to the same Locuslink ID were excluded from analysis. We had six subsets of data, normal and tumor subsets of the three data sets. We conducted the above correlation analysis 4,751,952 times for all combinations of data sets and pathways. This is a computationally intensive and time-consuming process. The correlation coefficient results were saved for computing coherence indicators of pathways and random gene sets.

The significance of coherence indicators was tested as follows. Let $G_K(D)$ denote the reference gene set. For a given data set and a pathway, we had n genes. We first calculated $N = n \times (n - 1)/2$ gene pairs in the pathway. We then computed the coherence indicator r_0 that was the estimate of the true ratio r of significantly correlated gene pairs in the pathway. The 95% confidence interval for the true ratio r was

$$[r_0 - 1.96 \times s/\mathrm{sqrt}(N), r_0 + 1.96 \times s/\mathrm{sqrt}(N)],$$

where $s = \mathrm{sqrt}(r \times (1 - r))$. However, this confidence interval was not useful because the ratio r was unknown. Since the ratio r was between 0 and 1, the upper bound of s was 0.5. Replacing s by 0.5, we obtained a wider confidence interval,

$$[r_0 - 0.98/\mathrm{sqrt}(N), r_0 + 0.98/\mathrm{sqrt}(N)].$$

To evaluate the significance of the coherence indicator r_0, we randomly selected n genes in the reference set $G_K(D)$ and computed the coherence indicator for this random gene set based on the same data subset used to compute r_0. We repeated this procedure 1000 times and obtained 1000 coherence indicators for the random gene sets. Since the true value of the coherence indicator was unknown, we defined the p value for r_0 as the frequency of the coherence indicators from random gene sets being greater than the left end of the confidence interval ($r_0 - 0.98/\mathrm{sqrt}(N)$). The left end of the confidence interval imposed more stringent selection of the significant coherent pathways. The fewer the coherence indicators of the random gene sets that were greater than the left end of the confidence interval, the more significant the coherence indicator r_0 became.

We have provided an Splus package to demonstrate our approach. The package contains a demo, more than 20 Splus functions, and some data files. It can be down loaded from ftp://ftp1.nci.nih.gov/pub/LeeLab/pathway.

Acknowledgment

We are grateful for having access to the publicly available gene expression data used in this analysis.

References

[1] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson Jr., L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, L.M. Staudt, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511.

[2] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.

[3] A. Soukas, P. Cohen, N.D. Socci, J.M. Friedman, Leptin-specific patterns of gene expression in white adipose tissue, Genes Dev. 14 (2000) 963–980.

[4] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, Proc. Natl. Acad. Sci. USA 96 (1999) 2907–2912.

[5] M.P. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr., D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Natl. Acad. Sci. USA 97 (2000) 262–267.

[6] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med. 7 (2001) 673–679.

[7] J.L. DeRisi, V.R. Iyer, P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (1997) 680–686.

[8] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, R.W. Davis, A genome-wide transcriptional analysis of the mitotic cell cycle, Mol. Cell 2 (1998) 65–73.

[9] T.G. Graeber, D. Eisenberg, Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles, Nat. Genet. 29 (2001) 295–300.

[10] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. USA 96 (1999) 6745–6750.

[11] D.A. Notterman, U. Alon, A.J. Sierk, A.J. Levine, Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays, Cancer Res. 61 (2001) 3124–3130.

[12] X. Chen, S.T. Cheung, S. So, S.T. Fan, C. Barry, J. Higgins, K.M. Lai, J. Ji, S. Dudoit, I.O. Ng, R.M. Van De, D. Botstein, P.O. Brown, Gene expression patterns in human liver cancers, Mol. Biol. Cell 13 (2002) 1929–1939.